International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

Heart Disease Patients: A Prediction Analysis Using Machine Learning Models

Mithun Das[1], Dr. Rajesh Das[2], Tapas Roy[3], Suman Acharjya[4]
[1]PhD Scholar Department of Library and Information Science, The University of Burdwan, West Bengal, India. http://orcid.org/0009-0007-6482-6425, 02mithun02@gmail.com
[2]Assistant Professor Dept. of Library and Information Science, The University of Burdwan, West Bengal, India. http://orcid.org/0000-0001-5349-6589, rajeshdas99@gmail.com
[3]PhD Scholar Department of Library and Information Science, The University of Burdwan, West Bengal, India. https://orcid.org/0009-0003-0610-5057, tapasroy135013@gmail.com
[4]MLIS Department of Library and Information Science The University of Burdwan, West Bengal, India. smnacharjya@gmail.com

Abstract
Cardiovascular disease is the number one killer disease worldwide that causes up to a third of deaths recorded every year. In 2022, it has claimed 19.8 million life in the Eastern Europe alone with mortality rates being equally or higher in the United States, Southeast Asia, and India. To minimize these figures, early recognition is important, and machine learning (ML) can facilitate it using abundant amounts of health data. In this paper, the researcher investigates how to predict heart disease using five ML models and a data set of 70,000 registered patients. The data had 13 clinical features and lifestyle elements like age, cholesterol, blood pressure, glucose, smoking, and alcohol consumption. The dataset was prepared after data pre-processing that is, feature selection, and normalization prior to the training and test of the model on a 70:30 data split. Gradient Boosting model also produced the highest performance accuracy of 73.48%, precision of 0.7637, recall of 0.690 and the F1-score of 0.7249. These were supported by results of confusion matrix that saw high values of true positive and true negative. This paper has shown that ML models, especially Gradient Boosting model, can greatly assist in the early detection of heart diseases in patients, which suggests their use in the clinical diagnostics area provided that the models are further modified and tested.

Keywords: *Heart disease, Prediction analysis, Machine learning model, Cosine similarity, Logistic Regression, Decision Tree classifier*

1. INTRODUCTION

Heart disease is one of the most serious killers that strikes individuals in the world because it is the reason behind around 33% of deaths worldwide every year. It affects the performance of the heart to play its usual role, which leads to the development of conditions such as heart attack and heart failure. A World Health Organisation report indicates that heart disease ranked as the leading cause of death in 2022, with 19.8 million fatalities in 100,000 people in Eastern Europe. Every 24 out of 100 people in the United States die from heart disease. The yearly number of deaths reaches 3.9 to 4.0 million in Southeast Asia.

According to the records of 2016, 28 per cent of the people passed away in India **(www.who.int)**. Hope lies in the fact that it is possible to improve machine learning in enhancing the accuracy of diagnosis because of the capability to process a large volume of health-related data. The research utilised a dataset of 70,000 patient records and 13 clinical and lifestyle options. These features include age, gender, blood pressure, cholesterol, level of glucose and lifestyle behaviours such as smoking and alcohol. The objective is to predict the presence and extent of heart disease using various classification models. Pre-processing of the data, feature selection and normalisation were involved to estimate the data quality. It applied five machine learning algorithms: Logistic Regression, Random Forest, Decision Tree, XGBoost and Gradient Boosting. The study aims to create a strong model that can facilitate the prediction of heart disease at an early stage to assist clinicians.

2. LITERATURE REVIEW

Mohan, Thirumalal & Srivastava (2019) noted that recent advances in machine learning have greatly improved the detection and forecasting of cardiovascular diseases. Various classification algorithms, such as support vector machines, decision trees, and logistic regression, have been employed to examine heart-

related concerns. Random forest ensembles have demonstrated effectiveness due to their ability to handle and integrate multiple features. Emerging research utilising health data from Internet of Things (IoT) devices has further enhanced model accuracy. Integrating diverse machine learning methods into hybrid systems has shown encouraging results and attracted considerable interest. However, many proposed models have not reached high generalizability due to limitations related to feature availability or dataset accessibility. More research is needed to analyse real-world data to create better and simpler models.

Shah, Patel & Bharti (2020) pointed out that heart diseases remain a major global cause of death despite available evidence. Efforts have been intensified to recognise early warning symptoms. The role of data mining techniques in healthcare diagnosis is expanding, allowing the analysis of intricate patient data. Numerous supervised learning techniques, such as naive bayes, decision trees, K-nearest neighbours, and random forests, have been used to forecast cardiovascular issues. The Cleveland dataset from the UCI repository is commonly used, often focusing on specific features. Among tested models, K-nearest neighbours frequently provide the most accurate results. These sophisticated data tools support healthcare professionals in making better-informed decisions regarding heart disease outcomes.

Bhatt, Patel, Ghetia & Mazzeo (2023) highlighted that addressing cardiovascular diseases involves timely, suitable medical interventions. Historical data trends show that machine learning techniques are increasingly applied to datasets of heart diseases, helping reduce misdiagnoses. Random forest, decision tree, multilayer perceptron, and XGBoost are methods for vast medical datasets. Improvements such as K-model clustering with Huang initialisation have shown potential for enhancing classification accuracy. Grid search CV (Cross-Validation) is used to tune model parameters, significantly boosting performance. Multilayer perceptron combined with cross-validation delivered the highest accuracy, making it a strong candidate for predictive health analysis.

Ramesh, Lihore, Poongodi, Simaiya, Kaur, & Hamdi (2022) emphasised that machine learning will be heavily relied upon for accurate disease diagnosis and treatment strategies due to the complex nature of healthcare data. Many researchers use publicly available datasets like the UCI heart disease repository to evaluate classification model performance. Effective feature selection is key to improving model accuracy. Isolation forest has helped enhance feature quality during classification pre-processing. Commonly used algorithms include KNN, Naive Bayes, SVM, Logistic Regression, Decision Tree, and Random Forest. Among them, KNN often outperforms when parameters are fine-tuned.

Dwivedi (2018) stated that heart diseases continue to be a global health issue, primarily driven by lifestyle risks and lack of awareness. Machine learning tools often promise to derive heart condition insights from extensive datasets. Many studies have compared algorithms using indicators such as accuracy, sensitivity, specificity, and ROC curves. The findings suggest that logistic regression is an effective model, with about 85% accuracy, offering solid performance in terms of sensitivity and specificity. The study concludes that the machine learning model is vital in timely disease identification and medical decision-making.

3. OBJECTIVES:

In general term, the objective of this study is to prediction analysis of heart disease of patients. The specifics abjectives are below:

I) To predict the severity of heart disease patients including heart attack.

iii) To compare algorithm performance across models

4. STATEMENT OF THE PROBLEMS:

Cardiovascular diseases (CVDS) are still the single greatest cause of mortality throughout the world, highlighting the vital importance of early detection and treatment to prevent them. Use of data-informed strategies, especially machine learning, would be a feasible means of improving the accuracy of diagnostics and increasing the effectiveness of timely intervention. This study uses a comprehensive data set of 70000 individual client health records and 13 clinical/lifestyle attributes to develop prediction models for cardiovascular disease. Such characteristics include demographic information (e.g., age and gender), physiological data (e.g., blood pressure, height, and weight), and lifestyle guides (e.g., smoking habits, alcohol consumption, and physical activity). This research will aim to examine the relationship between these variables and the presence of heart disease, meaning that the target variable is binary, and build a sound model that can detect high-risk individuals. This work is supported by the ambition to contribute to developing scalable, data-informed tools, which can help clinical decision making and preventive healthcare.

5. METHODOLOGY:

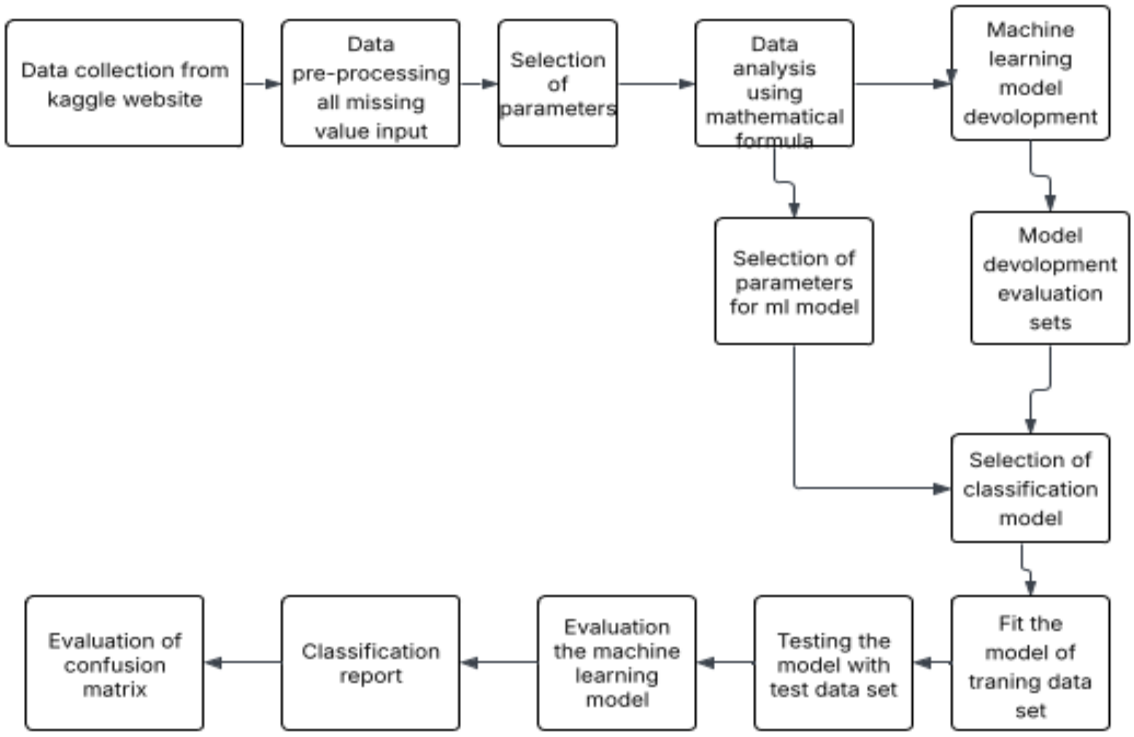


Figure-1: Framework for prediction analysis of heart disease

5.1. DATA COLLECTION:

In this research, we collected the "heart diseases" dataset CSV file (2.87 M.B) from www.kaggle.com, which was used for medical purposes to design our expected model. This CSV file contains 70000 (seventy thousand) patients and 13 (thirteen) attribute records. 13 (thirteen) attributes (details) are given below.

Attributes	Description
ID	1 to 70000
Age	Objective Feature | age | int (days)
Gender	Objective features.Male-1;Female-2
Height	Objective Feature | height | int (cm)
Weight	Objective Feature | weight | float (kg)
AP-hi(Systolic blood pressure)	Examination Features(At the time of heartbeats)
AL-10(Diastolic blood pressure)	Examination features (At the time of heart rest between beats).
Cholesterol	Subjective features: normal-1;above normal-2; well above normal-3.
Gluc	Subjective features: normal-1; above normal-2; well above normal-3.
Smoke	Subjective features.active-1.absent smoke-0
Alcohol Intake	Subjective features.active-1.absent alcol-0
Physical active	Subjective features.active-1.absent alcol-0
Target variable	“present cardio”-1, “absent cardio”-0.

Table-1: “Heart disease” dataset with all attributes

5.2. DATA PRE-PROCESSING:

Data processing is converting raw data to a clear data format. Due to the presence of raw data. It can not be an essential part of data mining techniques. So, data processing helps reduce data complexity during data analysis. Testing data quality and training for performance are the most important steps before applying machine learning algorithms. We download the CSV (heart disease) file from Kaggle to write our research paper. It estimates 13 parameters by following various strategies. Such as i) data cleaning, ii)

data conversion, iii) settlement of absence values, iv) data normalisation, v) feature selection, and so on, depending on the nature of the dataset. It can be easily implemented in a machine learning algorithm through various processes.

5.3. SLECTION OF PARAMETERS:

We downloaded the "heart diseases" CSV file from Kaggle, which contained records of 70000 patients and 13 attributes. To conduct machine learning analysis, we conducted this research work with a total of 12 (twelve) attributes, excluding 'ID' attributes (cause 'ID' attributes 1,2,3.....70000). Among the 12 attributes, 11(eleven) attributes are independent variables and 1(one) is the dependent variable. The dependent variable is derived from the characteristics of the 11 independent variables, where 'absent cardio is taken as a numerical value of '1' and 'present cardio ' is taken as a numerical value of '0', which is an important part of our research.

5.4. SELECTION OF MACHINE LEARNING MODELS:

The main objective of the research article is to predict 'present cardio'or 'absent cardio' from the target variable of the 'heart diseases' dataset, which consists of 70000 patient records and 13 attributes. To expect the best accuracy, precision, F-score, and recall score. We used machine learning algorithms: Logistic Regression(LR), Random Forest, Decision Tree, XGboost and Gradient Boosting.

5.5 MACHINE LEARNING TOOL:

In this experiment, we use Python for its vast library support, data processing capabilities and strong machine learning features. We conduct each examination using Google Colab. It reduces the need for local installations and configuration. It also provides strong computational resources, including a GPU(graphics processing unit) and at least 12 GB of RAM. These features significantly increase the training and performance of the machine learning model. The study implemented five machine learning classifications on the "heart diseases" dataset using Python in a colab environment. To check the functionality of the recommended models, we evaluate their performance using accuracy, precision and a confusion matrix. We also compare the performance of those systems.

5.6. DATA ANALYSIS AND INTERPRETATION:

In this research, we convert and filter the original dataset into a new one, and 12 attributes are selected. After that, we divided the dataset into two independent parts: 'present cardio' and 'absent cardio', which depend on the characteristics of heart disease.

The dataset used in this study is provided in a CSV file and imported using the pandas library, which offers powerful data manipulation and analysis tools. We utilise the sci-kit learn library to develop and evaluate machine learning models, which supports various algorithms, including logistic regression, Random Forest, decision tree classifier, and Gradient Boosting. Additionally, XGBOOTS is employed as a high-performance boosting algorithm. These tools also generate visualisations to support model interpretation and performance comparison.

5.7. EXPLORATORY DATA ANALYSIS:

In this research, we used the pandas command to detect missing values in the transformed dataset.

The df.insa () function creates a data frame at the same dimension as df (data frame), where at each level the corresponding value is true (Nan) if it is missing and false if it is present.figure-1 shows that our dataset command has no missing numerical values.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

```
df.isna().sum()
```

	0
id	0
age	0
gender	0
height	0
weight	0
ap_hi	0
ap_lo	0
cholesterol	0
gluc	0
smoke	0
alco	0
active	0
cardio	0

Figure-2: The dataset contains no missing values

Statistical description:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	sı
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.00(
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.00(
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.00(
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.00(
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.00(
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.00(

Figure-3: Descriptive Statistics of the Dataset

Before applying a machine learning strategy, the data exploration phase/ exploration of data analysis (EDA), is an essential primary step. Throughout this stage, statistical methods are used better to understand the structure and characteristics of the dataset.

Cardio description:

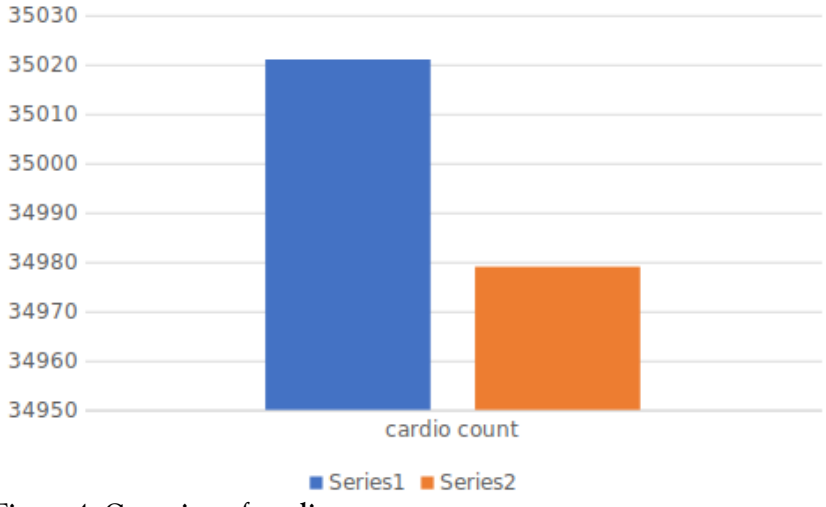


Figure-4: Counting of cardio

This is usually a serious content value like '0' and '1'. In research, '0' means 'absent cardio'; '1' means 'present cardio'. Figure 3 shows that the total number of 'absent cardio' is 35021 and 'present cardio' is 34979.

Selections of models

In this study, we have selected five classification models: logistic regression, random forest, decision tree, XG boost, and gradient boost.

Model development and evaluation sets:

```
[ ]  from sklearn.model_selection import train_test_split
     x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=1)
```

Figure 5: Overview describing the steps for selecting the machine learning model, emphasising dividing the dataset into the training and testing datasets.

Out of total 70000 records, 70%(49000) was allotted for training, and 30%(21000) used for examination. We worked with variables for independent attributes X and target attributes Y, which contain two user-defined classes. The data is split into X-train, Y-test, X-train, Y-train and Y-test based on the specific test size to separate the training and testing datasets.

Prediction analysis:

This study utilises a binary classification structure because it separates cardiovascular condition cases from cases without cardiovascular conditions. Various machine learning models were used to handle this classification issue. The employed models include logistic Regression, Random Forest, Decision Tree Classifier, XGBoost, and Gradient Boosting. The models have been built to perform efficient detection between two outcome categories. The dataset was split into training and testing groups for performance assessment and generalisation evaluation. The models received assessments according to their success in correctly categorising binary results.

```
from sklearn.linear_model import LogisticRegression
LR = LogisticRegression()
```

```
from sklearn.ensemble import RandomForestClassifier
RFC = RandomForestClassifier()
```

```
from sklearn.tree import DecisionTreeClassifier
DTC = DecisionTreeClassifier()
```

```
from xgboost import XGBClassifier
XGB = XGBClassifier()
```

```
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier()
```

Figure-6: Selected the Machine learning model.

After fitting five machine learning models to training and testing datasets, each model had its accuracy, as described below.

Machine learning models	Accuracy score
Logistic Regression	0.7183
Random Forest	0.7271
Decision Tree classifier	0.6366
XGBoost	0.7306
Gradient Boosting	0.7348

Table-2: ML model of accuracy results.

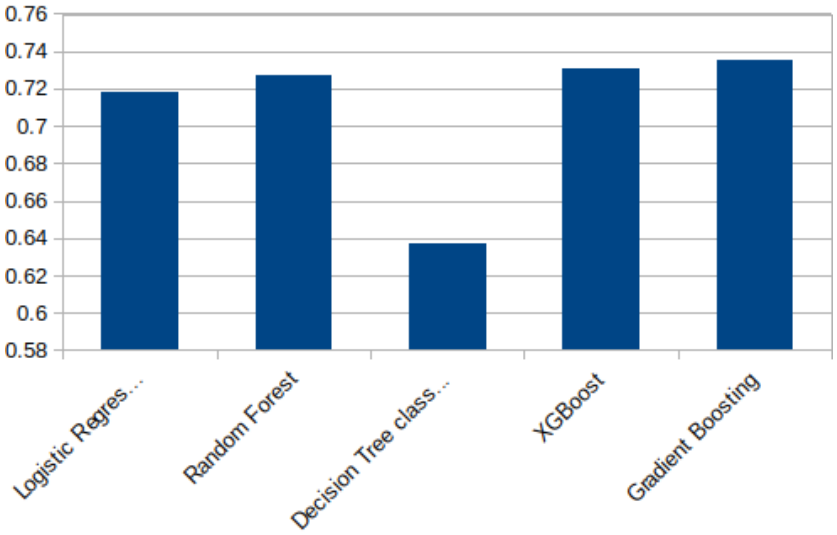


Figure-7: Column chart for ML models' accuracy scores

Figure 6 above shows that the gradient boosting model (0.7348) gave the best accuracy among the machine learning models in our research work, and the decision tree classifier (0.6366) showed the lowest accuracy. From all the accuracies, it is easy to see that the machine learning models we selected for the heart disease dataset performed very well.

Evaluation of the machine learning model:

Assessment of machine learning models is a primary step to deciding on practical deployment. The performance evaluation enables researchers to determine whether the model should be utilised as is or needs further development. The analysis executes predictive models through Logistic Regression and Random Forest, including Decision Tree and XGBoost, followed by Gradient Boosting. The models focus on achieving high accuracy results for cardiovascular condition identification. The model performance evaluation depends on applying suitable metrics for assessment. Measurements for model performance typically rely on statistical tools that evaluate classification effectiveness. The classification report and the confusion matrix remain two primary methods for accomplishing such analysis. These tools give users access to precision measurements, recall values, F1-scores, and total prediction accuracies.

Classification report:

```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.71      0.78      0.74     10352
           1       0.76      0.69      0.73     10648

    accuracy                           0.73     21000
   macro avg       0.74      0.74      0.73     21000
weighted avg       0.74      0.73      0.73     21000
```

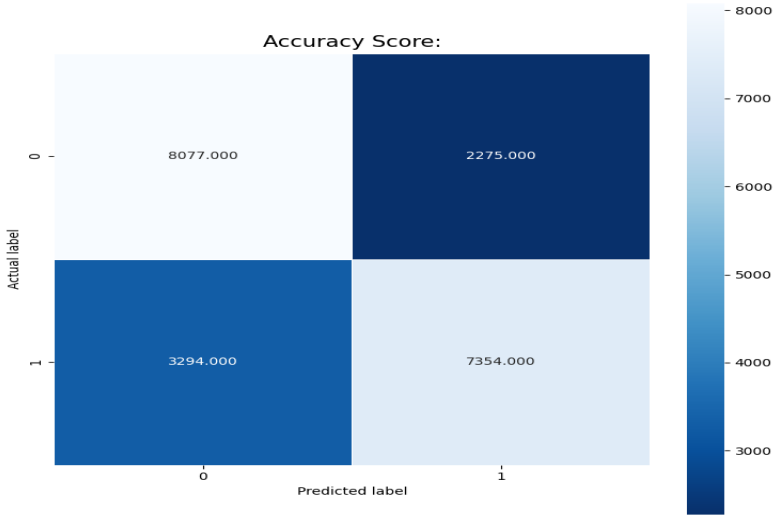
Figure-8: Classification report of the "y" test

The assessment report presents the evaluation of the models. The classification report is a popular evaluation method for any machine learning model. It comprises mainly 5 (five) columns and (N+3) rows. The first column is the class label name, followed by precision, recall, F-1 score and support. "N" rows are for "N" class labels (here binary class), and the other three rows are for accuracy, macro average and weighted average.

Model precision measures the accuracy of positive class predictions, and recall indicates its ability to find all cases of the positive class.

A unified measure of reliability for prediction models can be produced using the F1-score because it combines precision with recall. The model's overall precision is accurate, which measures the correct output rate of its responses. The model evaluates predicted classes between **'cardio'** and **'absent cardio'** with differing metric measurement values. The primary performance indicator, f1-score, reaches a value of **0.73**. The model demonstrates 73% precision in predicting the correct outcomes. The present score indicates that the classification functionality performs well. The model reflects enough performance to serve as a proper method for cardiovascular prediction. The model demonstrates satisfactory potential for future utilisation in applications and analysis.

Confusion matrix:



Confusion matrix and evaluation methods that-is widely used for machine learning models. A confusion matrix is an N*N table (where "N" is the number of classes) that contains the number of correct and incorrect predictions of the classification model. The confusion matrix includes four categories- True positive (TP), True negative(TN), False positive(FP) and False negative (FN).
True positive means that the model's predictive and real values are positive.
True negative means that the predictive and real values are positive.
False positive that is predictive is positive, and the real value is negative.
False negative means the predictive value is negative and the real value is positive.
Suppose the model gives an accuracy score above 70%. In that case, true positive and true negative values are very close, and the differences between false positive and false negative are minimal. Our confusion matrix shows that true negative and true positive values are 8077 and 7354, respectively. On the other hand, false positives and negatives are also very close, showing 2275 and 3294, which have very little difference. We can also model the classification report using TN, TP, FN, and FP values.
Accuracy=(TP+TN)/(TP+TN+FP+FN)=(7354+8077)/ (7354+8077+2275+3294)=15431/21000=0.734
Precision= TP/(TP+FP) = 7354/7354+2275 = 7354/9629 =0.7637
Recall=TP/(TP+FN) = 7354/(7354+3294) = 7354/10648 = 0.690
F-1 score = (2* precision*recall)/(precision+recall) =(2*0.7637*0.690)/(0.7637+0.690)
=1.053906/1.4537 = 0.7249
So, we have proved that the classification report is deeply related to a confusion matrix.

RESULTS

Five machine learning models were trained on the prediction of the dataset, split into its 70 per cent training and 30 per cent testing sets. Gradient boosting performed best as compared to other models, with its highest accuracy at 73.48% and Decision tree with the lowest accuracy at 63.66%. Large values of true positives (7354) and true negatives (8077) were observed in the confusion matrix, and there were relatively low values of false positives. The precision and recall of the best model were obtained as 0.7637 and 0.690, respectively. A balance of the two, F1-score, amounted to 0.7249, confirming the model's reliability. There is strong congruency between the classification report and the confusion matrix, which showed similar results. Visual analyses were also used to contribute to model performance comparisons. Generally, the Gradient Boosting model had the highest potential for accurate heart disease prediction.

CONCLUSION

This research shows that machine learning models can predict cardiovascular conditions based on clinical and lifestyle-oriented data. The comparative analysis implies that Gradient Boosting is the most precise and trusted model. The performance measure evaluation, including accuracy, precision, recall, and F1-score, confirmed the model's applicability in medical diagnostics. The confusion matrix analysis also supported the cross-validation of the results. The results indicate that the data-driven instrument may help medical workers find individuals at risk at an early stage. Its capabilities on mass data facilitate its practical usage. This method can be applied in clinical settings, subject to additional fine-tuning and validation. The following steps can be developing and increasing features and generalising the model.

REFERENCES

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/access.2019.2923707

[2] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science, 1(6). https://doi.org/10.1007/s42979-020-00365-y

[3] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088

[4] Tr, R., Lilhore, U. K., M, P., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. Malaysian. Journal of Computer Science, 132–148. https://doi.org/10.22452/mjcs.sp2022no1.10

[5] Dwivedi, A. K. (2016). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10), 685–693. https://doi.org/10.1007/s005 21-016-2604-1

[6] Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 128, 102289. https://doi.org/10.1016/j.artmed.2022.102289

[7]Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intelligence in Medicine, 16(1), 25–50. https://doi.org/10.1016/s0933-3657(98)00063-3

[8]Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088

[9] Amal, S., Safarnejad, L., Omiye, J. A., Ghanzouri, I., Cabot, J. H., & Ross, E. G. (2022). Use of Multi-Modal data and machine learning to improve cardiovascular disease care. Frontiers in Cardiovascular Medicine, 9. https://doi.org/10.3389/fcvm.2022.840262

[10] Purushottam, N., Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. Procedia Computer Science, 85, 962–969. https://doi.org/10.1016/j.procs.2016.05.288